doi:10.19920/j.cnki.jmsc.2024.09.004

基于离散数据流分割算法的预测方法①

孙丽君, 李方方*, 胡祥培
(大连理工大学经济管理学院, 大连 116024)

摘要: 离散数据流具有不确定性高、演变趋势多变以及多极值性等特征. 针对这些特征带来的预测难以兼顾准确性和实时性的问题,提出了基于离散数据流分割算法的预测方法. 该方法融合了短期趋势提取与分割点在线自适应检测:预测前基于非参数回归短期预测模型和改进的趋势提取算法获取短期(预测日)趋势,提前挖掘、分析预测日的短期趋势规律,以用于在线预测;预测时主要针对在线的数据流,基于假设检验自适应地检测分割点,以解决分割点难以确定的问题,并基于短期趋势修正分割点处的预测模型,减少了预测模型对缓冲数据的依赖. 数值实验结果验证了本研究所提预测方法的有效性和可行性.

关键词: 离散数据流预测; 在线数据流分割; 非参数回归; 趋势提取
中图分类号: F272.1　　**文献标识码**: A　　**文章编号**: 1007-9807(2024)09-0048-14

0 引　言

离散动态过程是由离散事件驱动的动态过程[1,2],若该过程运行状态的数据被连续、实时地采集,则会形成离散数据流[3]. 该数据流无处不在,如公共服务设施中的交通流、零售店中的客流、机械电子等各种离散型生产加工过程的数据流等. 离散动态过程运行中面临各种可能的"失控"状态,如,交通拥挤、商品缺货、机器故障等[4,5]. 因此,如何基于实时在线的离散数据流有效预测离散动态过程的未来演化趋势[6],以提前对过程"失控"预警,实现智能化的过程管理,是智能过程管理决策需解决的关键问题之一. 然而,离散动态过程一般具有不确定性高、未来演变趋势多变、随机突发影响因素较多的特点,导致该过程对应的离散数据流具有不确定性高、演变趋势多变以及多极值性的特征. 图 1 展示的由"车辆到达"这一离散事件驱动的某加油站客流(单位时间间隔内到达的车辆数)曲线明显呈现了该特征,这会严重影响预测方法的性能. 因此,研究考虑离散数据流特征的预测方法,无论对于实现离散动态过程的智能化管理,还是对于预测理论的丰富都具有十分重要的意义.

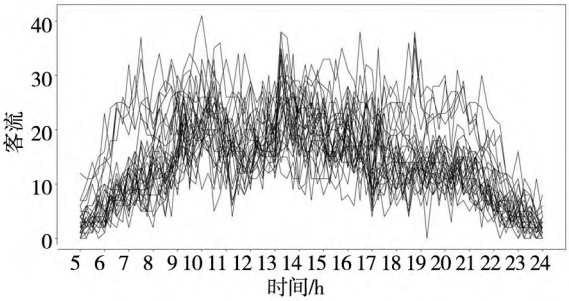

图 1　离散数据流示意图

Fig. 1 Graph of discrete data stream

本研究的离散数据流的实时预测方法主要针对单步预测. 该预测问题是一个时间序列预测问题[7]. 目前,时间序列预测方法主要有机器学习和时间序列分析两种[8,9]. 机器学习方法[10,11]有支持向量机(support vector machine, SVM)、循环神经网络(regression neural network, RNN)和长短期记忆(long short-term memory, LSTM)网络等. 时

① 收稿日期: 2020-09-02; 修订日期: 2022-06-10.
基金项目: 国家自然科学基金资助项目(71971037; 72371053).
通讯作者: 李方方(1992—), 女, 山东东营人, 博士生. Email: 1484029830@qq.com

间序列分析方法[12, 13]有指数平滑法(exponential smoothing,ES)、基于差分整合移动平均自回归模型(autoregressive integrated moving average, ARIMA)的改进时间序列预测模型等. 机器学习方法虽然预测精度较高,但是计算量大,计算速度相对较慢;而时间序列分析方法对于预测时段的跳跃式变化反应缓慢,只是单纯根据历史变化趋势预测未来. 因此,需要进一步研究兼顾实时性和精度的预测方法.

近年来,数据流在线分割算法[14-17]可以识别、表示出数据的规律性并实时检测到规律变化的时刻,也可以在合理的时间内对大数据集进行实时分析处理,可为离散数据流的实时预测提供方法支持[18, 19]. 目前该算法已广泛应用于股票价格预测[16, 17]、电力负荷预测[18, 19]等连续数据流领域. 该算法主要通过实时识别数据段的分割点,将数据流在线分割为一系列可变长度的数据段,采用回归函数拟合每个数据段,以最小化拟合值与观测值的误差,并用该回归函数进行预测. 该算法主要包含分割点检测和回归函数模型确定两部分. 根据回归函数的类型,数据流在线分割策略主要分为分段多项式表示[20]、分段线性表示[19, 21]等. 分段线性表示[22]虽然比二次或更高阶的多项式表示更容易确定,但是其只能表示有限的数据规律. 本研究为保证离散数据流预测的实时性和准确性,将基于分段多项式表示的数据流在线分割算法应用于离散数据流的实时预测中,既可以通过调整多项式的阶数控制拟合误差,也可以保持高压缩比和快速索引.

基于分段多项式表示的数据流分割算法包括自顶向下、自底向上和滑动窗口三类[22, 23]. 其中,自顶向下[16]算法根据斜率将时间序列片段递归分解为两个分量;自底向上算法[22]则根据误差阈值准则递归合并数据片段. 两种算法均基于整个数据集以离线方式分割数据,可以考虑数据的全局特性. 滑动窗口算法则是目前应用最广泛的数据流在线分割算法[15, 23, 24],该算法基于滑动窗口迭代扩展现有数据片段的边界,并构造片段表示方法以检测分割点,可以捕捉数据的在线特性. 该类研究主要分为两大类. 一类是不需要全局视角,仅仅基于实时到达数据以及事件发生的概率关系,实时检测分割点. 比如,文献[25]基于分离距离确定相邻窗口之间是否有足够的差异,以确定行为突然变化点;文献[26]通过不断扩展窗口并计算不同窗口下的人体活动概率密度函数,基于概率密度函数的最大值检测人体活动分割点. 另一类是融合自底向上等离线算法的全局优势,实时检测分割点. 该类算法中应用最广泛的是 Keogh[27]于 2001 年首次提出的一种结合离线自底向上分割算法和滑动窗口算法的混合在线算法(sliding window and bottom-up,SWAB). 该在线算法简便、快捷,且兼具自底向上、滑动窗口算法的优势. 然而,该算法分割点检测结果显著依赖于人工设定的阈值. 针对该阈值设定问题,已有数据流分割算法研究有以下两类自适应的分割点检测方式:1)基于斜率是否发生变化,自适应判断是否将窗口移动到下一个数据点,比如修正 SWAB(modified SWAB,mSWAB)算法[28]、嵌入式 SWAB(embedded SWAB,emSWAB)算法[29];2)基于离线训练集离线确定阈值,比如,文献[16]、文献[30]基于离线训练集离线确定出不同类别股票下的价格分段阈值,并用于在线阶段.

综上所述,目前数据流在线分割算法的相关研究主要基于阈值或自适应方式确定分割点,并在检测到分割点后基于前期缓存数据确定下一个时段的回归预测模型[19]. 现有研究中的分割点检测方式虽可为本研究提供思路,但是离散数据流呈现的多极值性以及高变化性导致以上研究中的分割点检测方式不能被直接应用,具体原因如下:1)用于确定分割点的阈值变化明显,人工设定阈值的负担大、科学性差;2)斜率频繁变化,且在线实时到达的数据与全局的离线数据特征差异大,分割点自适应检测难度大;3)分割点前后的变化趋势可能具有明显差异,如果继续使用缓存区内的数据构建预测模型,会影响预测精度. 为了弥补以上三个缺点,本研究在采用分段多项式表示以及基于 SWAB 的在线分割算法实现数据流预测的基础上,做出以下两点改进.

1)为了减少分割点检测对阈值的依赖,本研究引入预测误差的假设检验实现分割点的自适应

检测.

2)为了提高分割点处预测模型的预测精度并减少其对缓存数据的依赖,本研究引入短期趋势提取方法获取短期(预测日)趋势,以在检测到分割点后,将分割点处对应的短期趋势作为初始预测模型. 该短期趋势提取方法包含两部分:首先,采用完全数据驱动的非参数回归预测模型进行离散数据流的短期(一天)预测,克服离散数据流的分布规律难以统一的难题,得到更接近现实的短期预测结果[31];其次,采用基于动态规划的改进趋势提取算法提取短期预测结果中关键的离散数据流变化趋势信息. 因此,本研究提取的短期趋势融合了短期预测反映的离散数据流周期性规律以及趋势提取反映的趋势性规律,可以反映一段时间的离散数据流变化的方向,有利于提高分割点处的预测精度.

综上所述,本研究提出了基于离散数据流分割算法的预测方法,该方法融合了短期趋势提取与自适应的分割点检测:首先,在预测前,基于短期趋势提取方法提取预测日各个时段的离散数据流趋势以辅助在线实时预测;其次,针对在线到达的离散数据流,实时、自适应地检测分割点,并基于多项式回归预测模型和短期趋势进行预测.

1 基于离散数据流分割算法的预测方法的原理

基于 SWAB 在线分割算法的预测,采用多项式分段拟合[18]分割数据流,对数据流进行趋势偏移分析,实时检测出数据流趋势分割点(简称分割点);分割点在数据流趋势变化显著时,将未来数据流数据与现有数据分割开,相邻两分割点之间即形成一个趋势数据段,同一趋势段服从相同的回归预测模型. 因此,该算法的关键环节是确定分割点与预测样本数据. 目前该算法一般基于阈值[19]确定分割点和预测样本数据,其具体判别规则如下,其中,C、th_1、th_2分别表示预测误差累计值、阈值上限、阈值下限:1)当 $C \leqslant th_1$时,当前预测模型可继续用;2)当 C 介于 th_1、th_2之间时,当前预测模型可继续使用,但该时刻的数据需要放入缓冲区用于构建下一个时段的预测模型;3)当 C 超过 th_2时,说明该预测模型已不能达到良好的预测效果,此刻为分割点,需要利用缓冲区样本建立新的预测模型.

本研究在 SWAB 在线分割算法的基础上,根据离散数据流的特点,改进了分割点检测方法与预测模型,提出了离散数据流分割算法:1)基于预测误差的假设检验[32]自适应地判断实时到达的数据点是否是分割点;2)检测到分割点后,用基于历史数据提取的短期趋势辅助分割点处的初始预测模型. 基于该离散数据流分割算法的预测方法的原理如图 2 所示,具体步骤如下.

步骤 1 预测开始前,基于历史数据与短期趋势提取方法(参见第 4 节),获取预测日的短期(一天)趋势.

步骤 2 设初始预测时刻点为 T_1,数据段内的当前数据量 $N = 0$,基于步骤 1 确定的短期趋势,提取T_1 对应的趋势模型MF_0,并基于MF_0进行预测.

步骤 3 基于MF_0进行连续预测,直到某一时刻 T_i,时段 $[T_1, T_i)$ 内的数据量 N 超过阈值sn_1时,则基于该时段内的数据构建多项式回归预测模型(参见第 2 节)MF_1.

步骤 4 采用模型MF_1预测时刻 $T_i = T_i + \Delta T$(ΔT 表示采样时间间隔)的数据 $\widehat{Y}_{i+1}$,并在真实数据 Y_{i+1} 发生时,基于自适应分割点检测方法(参见第 3 节)判断该数据点是否是分割点. 如果该点不是分割点,说明其与MF_1的趋势仍然一致,则重复步骤 4;否则,令 $T_1 = T_i$,并返回到步骤 2.

综上所述,本研究提出的基于离散数据流分割算法的预测方法包含以下关键部分:1)多项式回归预测模型(第 2 节);2)自适应分割点检测方法(第 3 节);3)短期趋势提取方法(第 4 节).

2 多项式回归预测模型

设离散数据流的当前分段 Seg_i 的时间范围为$[T_1, T_i]$(T_1、T_i 分别表示分段开始时刻、当前

时刻). 当 Seg_i 中的数据量达到规定阈值后,本研究构建多项式回归预测模型进行离散数据流的预测. 具体地,本研究构建一元多项式回归模型(如式(1))拟合当前分段 Seg_i 内的数据 $Y = [Y_1, Y_2, \cdots, Y_i, \cdots, Y_N]^{\mathrm{T}}$,并利用该回归模型对离散数据流进行预测.

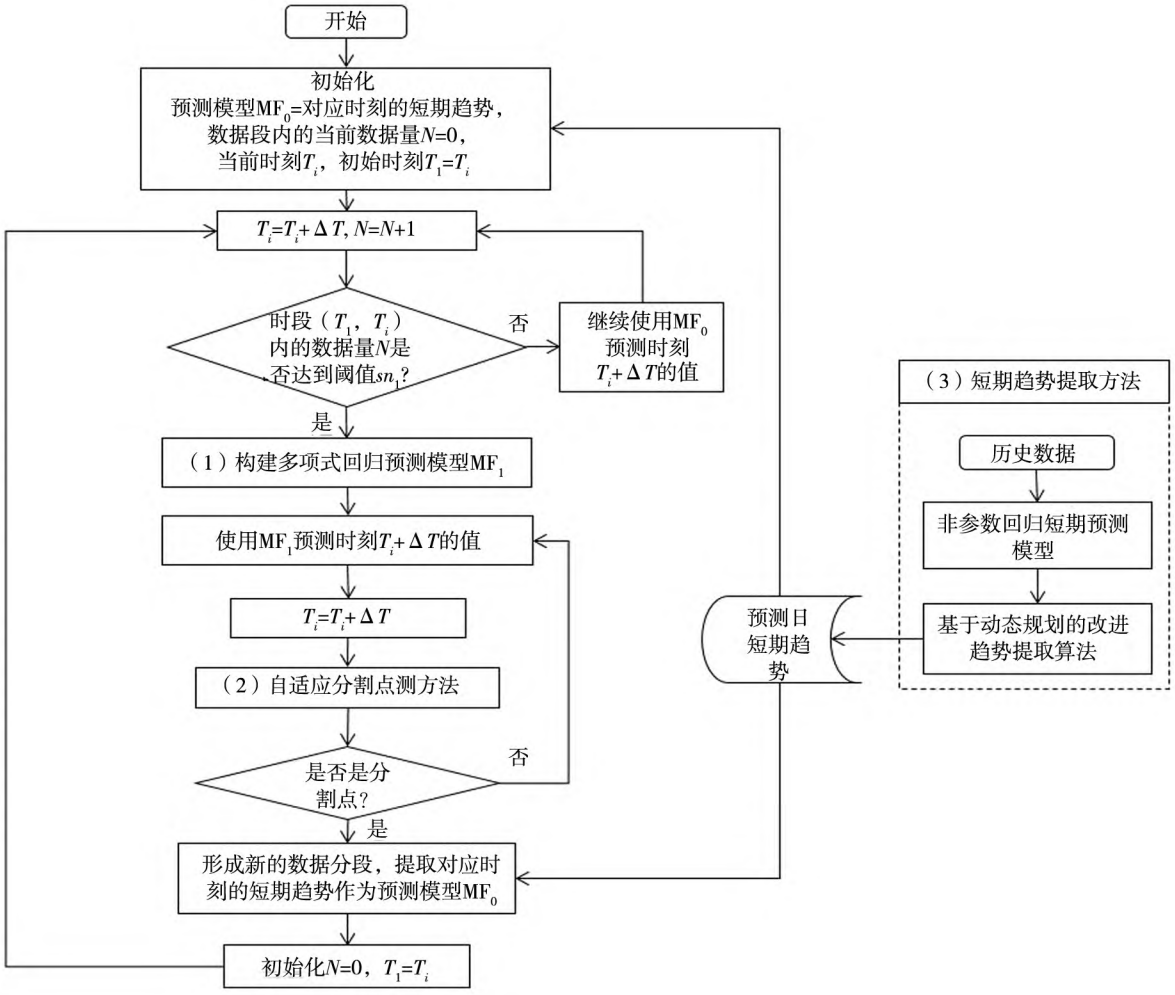

图 2　基于离散数据流分割算法的预测方法原理图

Fig. 2 Schematic of forecasting method based on on-line segmentation algorithm of discrete data stream

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1(T_i) + \cdots + \widehat{\beta}_m(T_i^m),\quad \forall i \in \{1,2,\cdots,N\} \tag{1}$$

其中 $\widehat{Y}_i$ 表示离散数据流的拟合值, T_i 表示时刻样本数据, N 表示分段内离散数据流的总数量, m 表示多项式的阶数.

回归模型的系数 $\widehat{\boldsymbol{\beta}} = [\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_m]^{\mathrm{T}}$ 采用标准最小二乘法[33]确定,即, $\widehat{\beta}$ 的取值如式(2). 基于 $\widehat{\boldsymbol{\beta}}$ 确定的多项式回归模型如式(3),基于该回归模型拟合段内数据的拟合误差 ε_{fit}^2 如式(4). 其中,回归函数的阶数 m 的确定采取 AIC(Akaike Information Criterion,赤池信息准则)准则(如式(5)),并令 $m \leqslant 2$ 以保证分段多项式拟合的速度能够应用于实时计算中.

$$\widehat{\boldsymbol{\beta}} = [\widehat{\beta}_0\ \widehat{\beta}_1\ \widehat{\beta}_2]^{\mathrm{T}} = (\boldsymbol{T}^{\mathrm{T}}\boldsymbol{T})^{-1}\boldsymbol{T}^{\mathrm{T}}\boldsymbol{Y} \tag{2}$$

$$\widehat{Y}_i = [\widehat{\beta}_0\ \widehat{\beta}_1\ \widehat{\beta}_2]\,[1\ (T_i - T_1)\ (T_i - T_1)^2]^{\mathrm{T}},\quad \forall i \in \{1,2,\cdots,N\} \tag{3}$$

$$\varepsilon_{fit}^2 = \frac{1}{N-m-1}\sum_{i=1}^{i=N}(Y_i - \widehat{Y}_i)^2 \tag{4}$$

$$\mathrm{AIC} = -2(N\log(RSS) + k) \tag{5}$$

其中 $(\boldsymbol{T})_{i,j} = T_i^{j-1}(1 \leqslant i \leqslant N,\ 1 \leqslant j \leqslant m)$, RSS 为残差平方和, k 为要估计参数的个数, $N - m - 1$ 表示 ε_{fit}^2 的自由度.

3 自适应分割点检测方法

设离散数据流的当前分段 Seg_i 对应的预测模型为MF_1(如式(3)). 该部分主要通过MF_1预测时刻 $T_i+\Delta T$ 的值 $\widehat{Y}_{i+1}$(如式(6)),在获取到该时刻的真实值 Y_{i+1} 后,基于该时刻的预测误差 ΔY_{i+1}(如式(7)) 进行 t 检验[32],以判断该时刻点是否是分割点,过程如下:首先,基于 t 检验确定用于检测分割点的阈值 $tv_{1,i}$; 然后,将$tv_{1,i}$ 与实际预测误差 ΔY_{i+1} 比较. 如果 $|\Delta Y_{i+1}| \leqslant tv_{1,i}$,说明该点不是分割点,未来数据流与预测模型 MF_1 的趋势仍然一致,可以继续采用模型 MF_1 进行预测; 否则,说明该点是分割点,未来数据流与 MF_1 的趋势不一致,需要构建新的预测模型进行预测.

$$\widehat{Y}_{i+1} = [\widehat{\beta}_0\ \widehat{\beta}_1\ \widehat{\beta}_2][1(T_i+\Delta T-T_1)\ (T_i+\Delta T-T_1)^2]^{\mathrm{T}} \tag{6}$$

$$\Delta Y_{i+1} = Y_{i+1} - \widehat{Y}_{i+1} \tag{7}$$

阈值 $tv_{1,i}$ 的推导如下所示.

设数据的真实值 Y_i 由其预测值 $\widehat{Y}_i$ 和随机误差 ε_i 构成,如式(8).

$$Y_i = \widehat{Y}_i + \varepsilon_i \tag{8}$$

其中预测随机误差 ε_i 服从高斯分布,$\mathrm{E}(\varepsilon_i)=0$,$\mathrm{E}(\varepsilon_i^2)=\sigma^2$.

基于式(6)、式(7)、式(8),预测误差 ΔY_{i+1} 如式(9).

$$\Delta Y_{i+1} = a_i^{\mathrm{T}}\boldsymbol{E}_i + \varepsilon_{i+1} \tag{9}$$

其中 $a_i=((\boldsymbol{T}_i^{\mathrm{T}}\boldsymbol{T}_i)^{-1}\boldsymbol{T}_i^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{Y}[\boldsymbol{1}(T_i+\Delta T-T_1)(T_i+\Delta T-T_1)]^{\mathrm{T}}$,$\boldsymbol{E}_i=[\varepsilon_1,\cdots,\varepsilon_i]^{\mathrm{T}}$,$\Delta Y_{i+1}$ 显然也服从高斯分布,方差如式(10).

$$Var(\Delta Y_{i+1}) = \delta^2(1+a_i^{\mathrm{T}}a_i) \tag{10}$$

真正的误差方差 δ^2 可以由分段$[T_1,T_i]$内所有数据的累积预测误差的方差 $\widehat{\delta}_i^2$ (式(11))估计,式(12)服从 t 检验,则预测误差 ΔY_{i+1} 的阈值 $tv_{1,i}$ 如式(13).

$$\widehat{\delta}_i^2 = \frac{\sum_{k=1}^{k=i}(Y_k-\widehat{Y}_k)^2}{i-m-1} \tag{11}$$

$$\frac{\Delta Y_{i+1}/\sqrt{\mathrm{Var}(\Delta Y_{i+1})}}{\sqrt{\widehat{\delta}_i^2/\delta^2}} = \frac{\Delta Y_{i+1}}{\widehat{\delta}_i\sqrt{1+a_i^{\mathrm{T}}a_i}} \tag{12}$$

$$tv_{1,i} = t_{1-\frac{\alpha}{2}}\widehat{\delta}_i\sqrt{1+a_i^{\mathrm{T}}a_i} \tag{13}$$

其中 $i-m-1$ 表示$\widehat{\delta}_i^2$ 的自由度,$t_{1-\frac{\alpha}{2}}$ 表示显著性水平为 α 的 t 检验值.

4 短期趋势提取方法

趋势分析是一种从大量过程数据中提取重要事件的有效工具,这也使其成为数据挖掘领域的重要工具[32, 34]. 趋势分析有两个主要任务:1)趋势提取,即将数据分割成一系列连续不重叠的片段(也叫趋势),片段的开始和结束时刻需要唯一地确定,以表示状态的变化; 2)趋势分析,即为每个片段分配基元(一种字母符号),用以描述片段内的时间变化特征. 为保证本研究提取的趋势可以作为预测模型,反映一段时间内的离散数据流变化方向,本研究的短期趋势提取方法首先采取非参数回归方法(4.1 节)进行离散数据流短期(一天)预测,其次采用趋势提取算法(4.2 节)将短期预测结果分割成一系列趋势,并确定趋势的多项式回归函数以作为基元. 最终,本研究的短期趋势提取方法可以将预测日离散数据流分割成若干数据段(趋势),每段有对应的时段范围(开始时刻、结束时刻)及相应的趋势拟合函数,一天的趋势集合即为预测日的短期趋势.

4.1 非参数回归短期预测模型

设 Y 为被解释变量,T 为解释变量,本研究所构建的非参数回归模型如式(14).

$$Y_i = k(T_i) + \varepsilon\varepsilon_i,\ \forall i \in \{1,2,\cdots,N\} \tag{14}$$

其中 Y_i 表示离散数据流的真实值,$k(\cdot)$ 是未知回归函数,$\varepsilon\varepsilon_i$ 为独立同分布的随机误差项,$\mathrm{E}(\varepsilon\varepsilon_i)=0$,$E(\varepsilon\varepsilon_i^2)=\sigma^2$. 当回归函数 $k(T_i)$ 的假设模型正确,模型中参数未知时,参数估计是最好的预测建模方法. 然而,实际中,$k(T_i)$ 的正确模型难以确定,数据驱动的非参数估计方法则可以仅从已知

的大量观测数据出发，实现对回归函数 $k(T_i)$ 的估计. 本研究采用常用的高斯核作为建立 $k(T_i)$ 的非参数回归估计的核函数.

4.2　基于动态规划的改进趋势提取算法

基于第 2 节的多项式回归预测模型，本研究采用基于多项式拟合的趋势提取算法，提取离散数据流短期预测数据中的趋势，即将该数据分割成若干数据段，并用多项式回归模型（式(3)）拟合每个数据段. 为最小化所有分段的拟合误差平方和，本研究在 Bo Zhou 等人[32]的研究基础上，基于动态规划思想确定趋势分割点. 因此，本研究趋势提取算法可转换为基于动态规划确定分割点的优化问题，如式 15 ~ 式 17 所示，根据数据 $Y_{1:t}=[Y_1,Y_2,\cdots,Y_i,\cdots,Y_t]$ 的最小拟合误差值 $F(t)$，依次递归地求解所有数据 $Y_{1:N_{all}}=[Y_1,Y_2,\cdots,Y_i,\cdots,Y_{N_{all}}]$ 的最小拟合成本 $F(N_{all})$，最终确定分割点集合 $S_{all}=\{s_i:s_0<s_1<\cdots<s_i<\cdots<s_M\}$ 的取值.

$$F(N_{all})=\min\left\{\sum_{i=1}^{M}\left[C(T_{(s_{i-1}+1):s_i})\right]\right\} \tag{15}$$

$$=\min\left\{\min\sum_{i=1}^{l}\left[C(T_{(s_{i-1}+1):s_i})\right]+C(T_{(t+1):N_{all}})\right\} \tag{16}$$

$$=\min\{F(t)+C(T_{(t+1):N_{all}})\} \tag{17}$$

其中分割点 s_0、s_M、s_l 分别取值为 0、N_{all}、t；成本函数 $C(T_{p:q})$ 为拟合误差平方和：$C(T_{p:q})=\frac{1}{q-p-m}\sum_{i=p}^{i=q}(Y_i-\widehat{Y}_i)^2$，$Y_i$ 表示实际值；$\widehat{Y}_i$ 表示拟合值；m 表示拟合多项式的阶数；$q-p-m$ 表示拟合多项式的自由度.

为提高各个数据段内数据点趋势的一致性以及趋势提取的速度，本研究在 Bo Zhou 等人[32]的研究基础上加入了以下两点：1）采取第 3 节的方法判断各个数据点与分段的趋势是否一致；2）采取剪枝算法[35,36]在保证成本全局最优的同时提高动态规划分割效率. 剪枝算法是假设当一个观测序列中增加一个分割点时，其成本函数取值会降低. 即，假设存在一个常数 K，对于所有的 $t<s_i<N_{all}$ 满足式(18)，则如果式(19) 能满足，t 不可能是 N_{all} 之前最后一个最优分割点.

$$C(T_{(t+1):s_i})+C(T_{(s_i+1):N_{all}})+K\leqslant C(T_{(t+1):N_{all}}) \tag{18}$$

$$F(t)+C(T_{(t+1):s_i})+K\geqslant F(s_i) \tag{19}$$

5　实例分析与数值实验

本研究的离散数据流是通过连续、实时采集离散动态过程的运行状态数据而形成的. 其中，运行状态可以通过单位时间内离散事件发生的次数反映. 本研究以大连市某加油站 2020 年 6 月 1 日至 2021 年 5 月 31 日的车辆到达数量，即客流数据为例进行实例分析，合计超 30 万条车辆到达记录. 数据采样时间间隔为 15min，总长度为35 040. 本研究首先采取标准的小波去噪方法对数据进行去噪预处理. 图 3 显示了部分预处理后的加油站日常客流样本，横坐标表示时刻点，两个时刻点之间的时间间隔为 15 min，纵坐标表示客流. 本研究将这些数据分为训练数据和测试数据，训练数据占全部数据的 70%，测试数据占 30%.

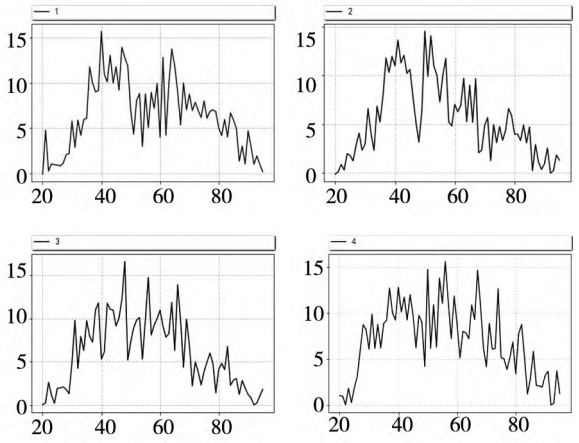

图 3　加油站客流样本数据示意图

Fig. 3 Graph of the sample data of customer flows in a gas station

本研究基于案例数据，统计了加油站每天的车辆到达情况，并通过 K-S 非参数检验[37]验证客流数据是否服从正态分布. K-S 检验是比较样本频率分布是否符合理论分布的一种假设检验方

法,其原假设 H0 为数据符合理论分布. 本研究基于 python 自带 K-S 检验程序包 scipy. stats. kstest 进行高斯性检验,返回 p 值,如果 $p>0.05$,接受 H0;反之,则拒绝 H0. 365 天的检验结果见图 4,横坐标表示日期,纵坐标表示 K-S 检验的 p 值. 其中,由于加油站 0 时至 5 时几乎没有车辆到达,客流数据分布检验中未体现这段时间.

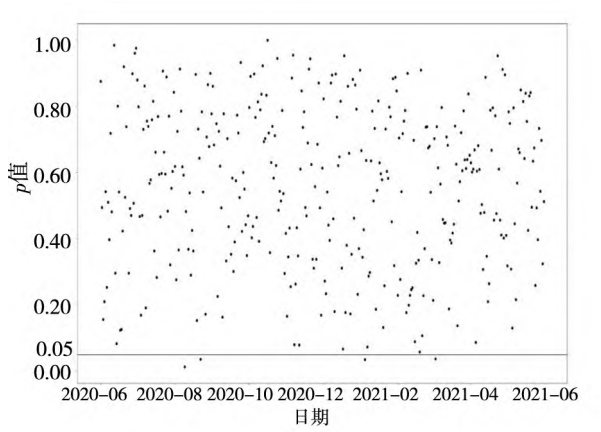

图 4 K-S 检验结果

Fig. 4 Results of K-S test

如图 4 所示,只有少数日期的客流数据不服从正态分布,占比约 3%. 其中,不服从正态分布的日期下的加油站主要是因为加油机系统故障维护、极端雷雨天气等原因造成的对应时段关闭,不符合正常的车辆到达规律,因此,这些日期可以不考虑.

根据 K-S 检验结果,有 354 天的加油站客流服从不同均值、方差下的正态分布且每天的客户到达情况均是相互独立的. 设每一天的加油站客流为随机变量 X_i,则服从正态分布的 354 天的加油站客流数据可以表示为随机变量集合$\{X_1, X_2, \cdots, X_{354}\}$. 根据李雅普诺夫定理(任意分布的 n 个随机变量 X,当 X 互相独立且 n 足够大时,n 个随机变量形成的总和随机变量服从正态分布,其中,$n>30$ 时,可称为大子样样本),本研究基于 354 天的加油站客流数据形成的总和随机变量近似服从正态分布. 满足以上正态分布条件的同时,加油站客户到达情况受足够多的随机因素影响,因此,加油站客流数据及其误差服从正态分布[38].

在评价预测模型性能时,本研究主要考虑绝对误差,选取的评价指标有平均绝对误差(mean absolute error, MAE)(式(20))、均方误差(mean square error, MSE)(式(21))和在线运行时间(Rt).

$$MAE = \frac{1}{N}\sum_{i=0}^{i=N-1}|Y_i - \widehat{Y}_i| \tag{20}$$

$$MSE = \frac{1}{N}\sum_{i=0}^{i=N-1}(Y_i - \widehat{Y}_i)^2 \tag{21}$$

在 MAE、MSE 的基础上,本研究进一步采用 Hansen 等[39]提出的模型信度检验(model confidence set, MCS)进行统计检验,以证明本研究所提预测模型与已有模型相比,预测效果具有显著优势[40]. MCS 方法不指定基础预测模型,检验结果的稳健性较强. MCS 方法的基本原理是检验所有预测模型构成的初始集合 M_0 中任意两个模型的预测效果是否存在差异,原假设 H_0(式(22))表示 M_0 中任意两个模型的预测效果相同[13].

$$H_{0,M}: E(d_{\varphi,l,l^*}) = 0 \ (l, l^* \in M, M \subset M_0) \tag{22}$$

其中 d_{φ,l,l^*} 表示M_0 中任意两个模型 l、l^* 在损失函数 φ 下的相对损失值. 如果特定置信水平下,显著拒绝原假设 $H_{0,M}$,则剔除M_0 中表现较差的一个模型,一直持续这一过程,直到不再拒绝原假设,没有模型被剔除. 该检验中用到的统计量包括半二次方统计量(式(23)).

$$T_{SQ} = \max \frac{(\overline{d_{\varphi,l,l^*}})^2}{Var(\overline{d_{\varphi,l,l^*}})}(l, l^* \in M, M \subset M_0) \tag{23}$$

其中,$\overline{d_{\varphi,l,l^*}}$ 是d_{φ,l,l^*} 的平均值,T_{SQ} 对应的 p 值通过自举法确定,p 值越大,表明模型的预测精度越高. 本研究在显著性水平 0.01 下进行 MCS 检验,验证所提预测方法是否能显著改进预测性能.

为验证所提预测方法的有效性,本研究进行了两大类实验:1)第一类包含三个实验,用于验证所提离散数据流分割算法的有效性(5.1 节):①进行不同分割点检测方法的对比试验,验证了本研究分割点检测方法对于预测结果的优势;②获取短期预测结果、趋势提取结果,并进行不同趋

势提取算法的对比实验，验证了本研究所提趋势提取算法本身及其对于改进预测效果的优势；③进行分割点处不同预测模型构建方法的对比实验，验证了本研究提取的短期趋势对提高预测精度的效果；2）第二类进行不同预测方法的对比实验，用于验证本研究所提预测方法的预测效果（5.2 节）. 下文对两类实验进行详述.

5.1　离散数据流分割算法有效性验证分析

5.1.1　分割点检测方法的效果验证

为验证本研究所提自适应分割点检测的有效性，在使用同一短期预测结果的基础上，对比基于本研究所提分割点检测方法与基于阈值的分割点检测方法[19]的预测效果. 由于基于阈值的分割点检测方法受阈值上限 th_1、下限 th_2 的影响，本研究所提分割点检测方法也受多项式回归预测模型构建所需的初始数据量 sn_1 的影响. 为此，本研究对 th_1、th_2、sn_1 进行敏感性分析，以确定方法的最佳预测结果.

1）th_1、th_2 的敏感性分析

由于不同预测日基于相同取值的 th_1、th_2 进行预测的效果差异大，因此，本研究在每天工作开始前，基于短期预测结果进行敏感性分析：具体地，令阈值 th_1 由 5 变化到 40，步长为 5；在 th_1 一定时，令阈值 th_2 由（th_1 + 1）变化到（th_1 + 50）. 某一天的参数敏感性分析部分结果如图 5（a）、图 5（b）所示，图 5（a）表示 th_1 = 40 时，预测评价指标 *MSE* 随 th_2 变化的情况；图 5（*b*）中 *x* 轴表示 th_1，*y* 轴表示 th_2，*z* 轴表示 *MSE*，实心散点表示特定 th_1 下最小的 *MSE* 及其对应的 th_2.

2）sn_1 的敏感性分析

sn_1 对预测效果的影响在不同日期的差异较小，本研究基于所有测试集开展敏感性分析，以确定唯一、固定的值. 由于 sn_1 表示多项式回归预测模型构建所需的初始数据量，其最小取值为 2，且初始值不宜过大，因此，令 sn_1 由 2 变化到 6，步长为 1，所有测试日期的 *MSE* 随 sn_1 的变化情况如图 5（c）所示.

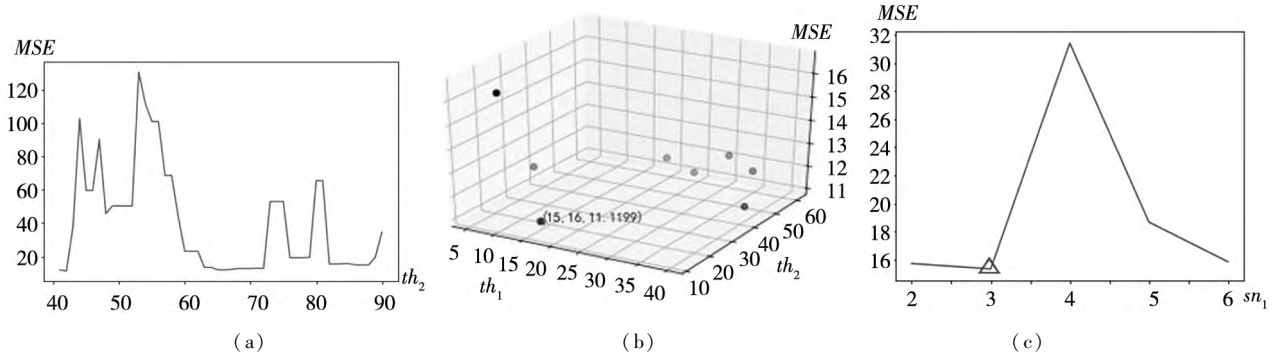

图 5　阈值的敏感性分析的结果示意图

Fig. 5 Graph of sensitivity analysis results of thresholds

图 5（*a*）显示，th_1 一定时，*MSE* 不会随 th_2 的变化呈现规律性变化；图 5（*b*）显示，点（15，16，11.119 9）对应最小的 *MSE* = 11.119 9，即当 th_1 = 15，th_2 = 16 时，该天基于该分割点检测方法的预测效果最佳. 图 5（*c*）显示，sn_1 = 3 时，*MSE* 取得最小值，为 15.361 2. 虽然 sn_1 = 6 时，*MSE* 的取值与 sn_1 = 3 时相差不大，但是 sn_1 越大，需要更多地依赖于短期趋势提取值，会忽视实时发生的现场情景，因此，本研究令 sn_1 = 3.

获取两种分割点检测方法的最佳预测结果后，对比两种方法的 *Rt*、*MAE*、*MSE*、以及 *MAE*、*MSE* 两种损失函数下的 MCS 检验结果，结果如表 1 所示.

表 1 显示基于本研究所提分割点检测方法的预测在均方误差（*MSE*）、平均绝对误差（*MAE*）两个评价指标上得到的预测误差值均小于基于阈值的数据流分割预测方法；本研究所提方法在 2 个损失函数下的 MCS 检验的 *p* 值均为 1，对比方法的 *p* 值均为 0，证明本研究所提方法具有显著的

预测效果改进;以上两点证明了本研究所提分割点检测方法对于预测的有效性. 尤其是,由于基于阈值的数据流分割预测方法需要在每天预测前,基于短期预测值确定最佳 th_1、th_2,这导致基于本研究所提分割点检测方法的训练时间远小于基于阈值的数据流分割预测方法. 虽然基于本研究所提方法的在线运行时间 *Rt* 比基于阈值的方法大,但是平均到每一天,仅为 0. 166 1 s,完全可以保障方法的在线实时性. 综上所述,本研究所提方法在 *MSE*、*MAE*、*Rt*、MCS 检验结果上的表现能够充分说明其在减少阈值依赖的同时,提高了预测的精度.

表 1 预测结果比较

Tabel 1 Comparison of forecasting results

评价指标	*MAE*	*MAE* 下 MCS 检验的 *p* 值	*MSE*	*MSE* 下 MCS 检验的 *p* 值	*Rt*/s	训练时间/s
本研究所提方法	**2. 998 9**	**1. 0**	**17. 610 2**	**1. 0**	17. 113 4	**5 098. 950 6**
对比方法	4. 051 2	0. 0	38. 119 9	0. 0	**11. 172 8**	7 543. 384 9

注: *p* 值越大模型精度越高,粗体表示在所有模型中最优的取值.

5. 1. 2 *改进趋势提取算法对客流预测的效果验证*

基于本研究所提的非参数回归短期预测模型确定预测日的客流值,图 6 中的虚线表示某日的预测客流曲线,实线表示真实客流曲线.

获取客流的短期预测值后,本研究采用基于动态规划的改进趋势提取算法提取各个时段的趋势. 为了验证该方法的有效性,本研究将其与基于自底向上和滑动窗口(SWAB)以及基于动态规划(DP)的两种趋势提取算法[32]比较:1)首先,为验证趋势对数据拟合的好坏,比较三种算法的 *MAE*、*MSE* 以及数据的分段总数量,结果如表 2 所示;2)其次,为验证趋势提取结果对客流预测的影响,本研究使用 *MAE*、*MSE*、*Rt*、*MCS* 检验的 *p* 值对三种趋势提取结果下的预测效果进行分析,结果如表 3 所示. 其中,三种趋势提取算法下的离线训练时间为非参数回归短期预测模型的训练时间,不需比较.

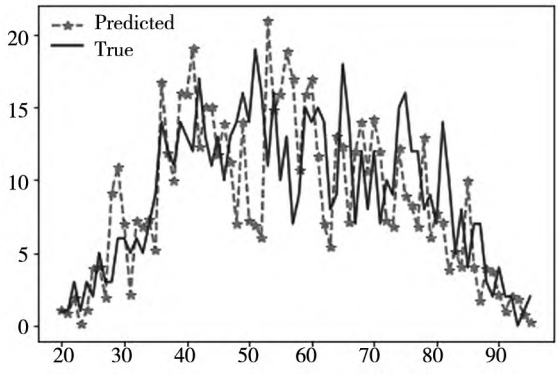

图 6 预测日的真实及预测的客流曲线

Fig. 6 The true and predicted customer flow curves

表 2 不同趋势提取算法的数据拟合结果

Tabel 2 Data fitting results of different trend extraction algorithms

评价指标	本研究所提趋势提取算法	基于 DP 的趋势提取算法	基于 SWAB 的趋势提取算法
MAE	**3. 762 51**	3. 908 71	3. 838 465
MSE	**28. 864 4**	31. 534 4	30. 461 4
分段总数	11	16	18

表 3 基于不同趋势提取算法的预测结果

Tabel 3 Forecasting results based on different trend extraction algorithms

评价指标	*MAE*	*MAE* 下 MCS 检验的 *p* 值	*MSE*	*MSE* 下 MCS 检验的 *p* 值	*Rt*/s
基于本研究所提趋势提取算法的预测方法	**2. 998 9**	**1. 0**	**17. 610 2**	**1. 0**	17. 113 4
基于 DP 趋势提取算法的预测方法	3. 090 5	0. 0	20. 141 5	0. 0	**14. 894 8**
基于 SWAB 趋势提取算法的预测方法	3. 006 6	**0. 333 3**	17. 873 1	0. 0	16. 715 8

表 2 显示，本研究所提趋势提取算法的 *MAE*、*MSE* 均小于基于 DP 和基于 SWAB 的趋势提取算法，且本研究所提趋势提取算法分段总数量最小，证明本研究所提趋势提取算法可以以最小的分段数、最高的预测精度实现对数据的拟合. 表 3 显示，在 *MAE*、*MSE* 两个评价指标上，基于本研究所提趋势提取算法的预测方法的预测误差值均小于基于 DP 和 SWAB 趋势提取算法的预测方法；且本研究所提方法在 2 个损失函数下的 p 值均为 1，证明本研究所提趋势提取算法相对于其余两种算法对预测性能有显著改进；基于SWAB算法的预测方法在 *MAE* 下 MCS 检验的 p 值为 0.333 3，证明该算法对预测性能的改进仅次于本研究所提算法. 综上所述，本研究所提趋势提取算法有较好的结果，基于该算法提取的某一天的短期趋势，即各个分段的时段范围（开始时刻、结束时刻）及相应的趋势拟合函数见表 4.

表 4　趋势提取结果

Tabel 4 Results of the trend extraction

时段范围	拟合函数	时段范围	拟合函数	时段范围	拟合函数
[20, 26)	$y = 0.006\,974x^2 - 0.160\,9x + 0.943\,5$	[35, 49)	$y = 8.461$	[66, 73)	$y = 0.398\,8x^2 - 55.02x + 1\,902$
[26, 29)	$y = -1.245x^2 + 69.04x - 951.8$	[49, 55)	$y = 0.338\,1x^2 - 34.07x + 864.2$	[73, 92)	$y = 0.027\,94x^2 - 4.887x + 215.8$
[29, 35)	$y = 9.218$	[55, 66)	$y = 8.487$	[92, 95)	$y = -0.817\,1x + 77.03$

5.1.3　*短期趋势提取对预测的效果验证*

为验证本研究提取的短期趋势对客流预测的效果，本研究使用指标 *MAE*、*MSE*、MCS 检验的 p 值、*Rt* 对比本研究所提方法与基于缓存区数据（对比方法 1）以及直接利用短期预测结果（对比方法 2）构建分割点处预测模型下的预测结果，如表 5 所示.

表 5　分割点处不同预测模型构建方法的预测结果

Tabel 5 Forecasting results of different forecasting models at segmentation points

评价指标	*MAE*	*MAE* 下 MCS 检验的 p 值	*MSE*	*MSE* 下 MCS 检验的 p 值	*Rt*/s	训练时间/s
本研究所提方法	**2.998 9**	**1.0**	**17.610 2**	**1.0**	17.113 4	**5 098.950 6**
对比方法 1	4.051 2	0.0	38.119 9	0.0	**11.172 8**	7 543.384 9
对比方法 2	3.151 2	0.0	20.183 8	0.0	15.836 7	**5 098.950 6**

表 5 显示，本研究所提方法在 *MAE*、*MSE*、两种损失函数下的 MCS 检验 p 值上的表现均是最优的，证明将本研究提取的短期趋势用于辅助客流预测，有利于提高预测的精度.

5.2　与常用预测方法的对比实验

为了验证本研究所提预测方法的有效性，对比其与常用的机器学习方法（SVM、DA-RNN（dual-stage attention-based recurrent neural network，双阶段的注意力机制循环神经网络模型））、线性回归方法以及时间序列分析方法（SARIMA（seasonal auto regressive integrated moving average，季节性自回归移动平均模型））的预测结果. 由于计算限制，本研究每天基于固定大小的滚动训练数据集，更新预测模型的参数. 其中，DA-RNN 结合了注意力机制和 seq2seq 结构（一种 encoder-decoder 结构），其关键参数包括分割窗口长度 T、梯度更新块大小 *batch_size*、学习率 η、编码器的隐藏状态大小 m 和解码器的隐藏状态大小 p. 为简便起见，本研究设 $m = p$. 为了确定各参数的取值，令 $T = \{2, 3, 4, 5, 10, 15\}$、$m = p = \{16, 32, 64, 128, 256\}$、*batch_size* $= \{64, 128\}$、$\eta = \{0.01, 0.02, 0.03, 0.05, 0.1, 0.12, 0.13, 0.15, 0.2, 0.3, 0.4, 0.5\}$，进行预测效果验证. 当 $T = 2$、$m = p =$ *batch_size* $= 128$、$\eta = 0.1$ 时，DA-RNN 取得最佳效

果. 其中,$T=2$、$m=p=batch\_size=128$ 时,不同 η 下的 *MSE* 取值见图 7. 此外,$SARIMA(p, d, q)(P, D, Q, s)$ 的季节性参数 s 为 1 天的数据采样点数,其余参数基于 AIC 准则和训练集迭代确定;SVM 选择常用的高斯径向基函数作为核函数.

本研究使用 *MAE*、*MSE*、MCS 检验的 p 值、*Rt*、训练时间对预测结果进行分析,结果如表 6 所示.

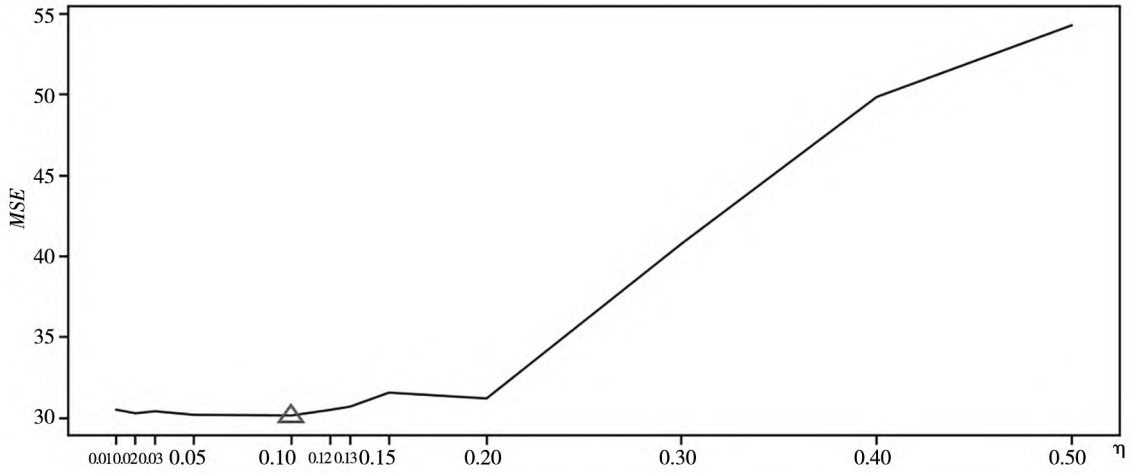

图 7　η 的敏感性分析的结果示意图

Fig. 7 Graph of sensitivity analysis results of η

表 6　预测结果比较

Tabel 6 Comparison of forecasting results

评价指标	*MAE*	*MAE* 下 MCS 检验的 p 值	*MSE*	*MSE* 下 MCS 检验的 p 值	*Rt*/s	训练时间/s
本研究所提方法	**2.998 9**	**1.0**	**17.610 2**	**1.0**	17.113 4	5 098.950 6
SVM	3.029 9	0.0	18.122 8	**0.333 3**	2.601 2	7 116.315 1
DA-RNN	4.080 2	0.0	30.152 8	0.0	0.984 4	368.364
线性回归	3.319 5	0.0	21.304 0	0.0	**0.028 3**	**106.858 5**
Arima	3.463 9	0.0	22.303 7	**0.333 3**	64.840 9	21 264.038 1

表 6 显示,在 *MSE*、*MAE* 两个评价指标上,本研究提出的预测方法得到的预测误差值是最小的;且在 2 个损失函数下 MCS 检验的 p 值均为 1,证明本研究预测方法相对于其余四种预测方法具有显著最强的预测能力,可以充分证明本研究预测方法的有效性. 从表 6 中也可以看出以下两点信息:1) DA-RNN 预测方法的预测误差在 *MSE*、*MAE*、MCS 检验的 p 值上的表现均是最差的,这是因为神经网络面对类似本研究的低维数据流时,大部分情况下比不过 SVM、线性回归等;2) SVM、SARIMA 在损失函数 *MSE* 下 MCS 检验的 p 值均为 0.333 3,证明这两种方法的预测能力仅次于本研究所提方法,对低维离散数据流有一定的适用性;但是 SVM 等机器学习算法以及 SARIMA 等时间序列分析方法由于其计算模型的复杂性,导致其训练时间过长,特别是 SARIMA 的在线运行时间过长,因此不利于在线实时的预测.

6　结束语

本研究提出了基于离散数据流分割算法的预测方法. 该方法的创新之处在于:1) 将基于预测误差的假设检验和短期趋势应用于离散数据流分割算法,自适应地确定分割点,并基于短期趋势修

正分割点处的预测模型，减少了预测模型对缓冲数据的依赖，解决了具有多极值性、高变化性的离散数据流分割点难以确定以及分割点前后趋势变化明显的问题，提高了预测的准确率；2）将非参数回归短期预测模型和基于动态规划的改进趋势提取算法相结合构建短期趋势提取方法，使得本研究提取的短期趋势融合了短期预测反映的离散数据流周期性规律以及趋势提取反映的趋势性规律，可以反映一段时间的离散数据流的变化方向，有利于提前挖掘和分析预测日离散数据流的趋势规律，克服了不确定性高、未来演变趋势多变的离散数据流趋势提取难的问题，也丰富了趋势提取方法的应用领域；3）使用完全由数据驱动的非参数回归方法为离散数据流建立短期预测模型，克服了离散数据流模式复杂多变导致的短期预测准确率低的缺陷，并丰富了非参数回归方法的应用领域.

本研究所提的预测方法可以广泛应用于由离散事件驱动的动态系统的在线实时监控中，可以在线实时实现系统流量的预测，以提前预警系统的潜在失控状态（如，交通拥挤、商品缺货、机器故障等），对离散动态系统的智能化调控具有重要的现实意义. 此外，该方法也可以用于其它具有多极值性、趋势多变性特征的单维时间序列数据流的预测问题，这也将是本研究未来的拓展方向.

参 考 文 献：

[1] Liu Y, Sun R, Jin S. A survey on data-driven process monitoring and diagnostic methods for variation reduction in multi-station assembly systems[J]. Assembly Automation, 2019, 39(4): 727 – 739.

[2] Zhao S, Shmaliy Y S, Ahn C K, et al. Probabilistic monitoring of correlated sensors for nonlinear processes in state space [J]. IEEE Transactions on Industrial Electronics, 2020, 67(3): 2294 – 2303.

[3] Zubaroğlu A, Atalay V. Data stream clustering: A review[J]. Artificial Intelligence Review, 2021, 54: 1201 – 1236.

[4] Topan E, Van Der Heijden M C. Operational level planning of a multi-item two-echelon spare parts inventory system with reactive and proactive interventions[J]. European Journal of Operational Research, 2020, 284(1): 164 – 175.

[5] Montoy R, Gonzalez C. A hidden markov model to detect on-shelf out-of-stocks using point-of-sale data[J]. Manufacturing & Service Operations Management, 2019, 4(21): 932 – 948.

[6] 刘玉敏, 周昊飞. 基于小波重构与 SVM-BPNN 的动态过程在线智能监控[J]. 系统工程理论与实践, 2016, 7(36): 1890 – 1897.
Liu Yumin, Zhou Haofei. Online intelligent monitoring for dynamic process based on wavelet reconstruction and SVM-BPNN [J]. Systems Engineering: Theory & Practice, 2016, 7(36): 1890 – 1897. (in Chinese)

[7] 梁志峰, 王　铮, 冯双磊, 等. 基于波动规律挖掘的风电功率超短期预测方法[J]. 电网技术, 2020, 44(11): 57 – 65.
Liang Zhifeng, Wang Zheng, Feng Shuanglei, et al. Ultra-short-term forecasting method of wind power based on fluctuation law mining[J]. Power System Technology, 2020, 44(11): 57 – 65. (in Chinese)

[8] Fildes R, Ma S, Kolassa S. Retail forecasting: Research and practice[J]. International Journal of Forecasting, 2019, 29 (1): 41 – 56.

[9] Xu S, Chan H K, Zhang T. Forecasting the demand of the aviation industry using hybrid time series sarima-svr approach [J]. Transportation Research Part E: Logistics and Transportation Review, 2019, 122: 169 – 180.

[10] Huber J, Stuckenschmidt H. Intraday shelf replenishment decision support for perishable goods[J]. International Journal of Production Economics, 2021, 231: 107828.

[11] Tang L, Zhao Y, Cabrera J, et al. Forecasting short-term passenger flow: An empirical study on shenzhen metro[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(10): 3613 – 3622.

[12] Tang J, Zuo A, Liu J, et al. Seasonal decomposition and combination model for short-term forecasting of subway ridership [J]. International Journal of Machine Learning and Cybernetics, 2022, 13: 145 – 162.

[13]陈 王, 魏 宇, 马 锋, 等. 高频视角下股市波动预测的新方法: HARFIMA 模型[J]. 管理科学学报, 2020, 23(11): 103 - 116.
Chen Wang, Wei Yu, Ma Feng, et al. A new method of stock market volatility forecasting in high-frequency per-spective: HARFIMA model[J]. Journal of Management Sciences in China, 2020, 23(11): 103 - 116. (in Chinese)

[14]Nascimento D C, Pimentel B, Souza R, et al. Dynamic time series smoothing for symbolic interval data applied to neuroscience[J]. Information Sciences, 2020, 517: 415 - 426.

[15]Qian H, Pan S J, Miao C. Weakly-supervised sensor-based activity segmentation and recognition via learning from distributions[J]. Artificial Intelligence, 2021, 292: 103429.

[16]Chen Y, Hao Y. A novel framework for stock trading signals forecasting[J]. Soft Computing, 2020, 24(16): 12111 - 12130.

[17]Tang H, Dong P, Shi Y. A new approach of integrating piecewise linear representation and weighted support vector machine for forecasting stock turning points[J]. Applied Soft Computing, 2019, 78: 685 - 696.

[18]Li Y, Wang T, Deng Z. An online reactive power-optimization strategy based on load-curve prediction and segmentation [J]. Applied Sciences, 2020, 10(3): 1145.

[19]唐聪岚, 卢继平, 谢应昭, 等. 基于改进数据流在线分割的超短期负荷预测[J]. 电网技术, 2014, 38(7): 2014 - 2020.
Tang Conglan, Lu Jiping, Xie Yingzhao, et al. Improved data stream on-line segmentation based ultra short-term load forecasting[J]. Power System Technology, 2014, 38(7): 2014 - 2020. (in Chinese)

[20]Charbonnier S, Becq G, Biot L. On-line segmentation algorithm for continuously monitored data in intensive care units[J]. IEEE Transactions on Biomedical Engineering, 2004, 51(3): 484 - 492.

[21]Kim J H, Hagiwara T. L1 optimal controller synthesis for sampled: Data systems via piecewise linear kernel approximation [J]. International Journal of Robust and Nonlinear Control, 2021, 31: 4933 - 4950.

[22]Feng G, Zhang L, Yang J, et al. Long-term prediction of time series using fuzzy cognitive maps[J]. Engineering Applications of Artificial Intelligence, 2021, 102: 104274.

[23]Bhattacharya D, Mukhoti J, Konar A. Learning regularity in an economic time-series for structure prediction[J]. Applied Soft Computing, 2019, 76: 31 - 44.

[24]Zhao H, Li G, Zhang H, et al. An improved algorithm for segmenting online time series with error bound guarantee[J]. International Journal of Machine Learning and Cybernetics, 2016, 7(3): 365 - 374.

[25]Aminikhanghahi S, Wang T, Cook D J. Real-time change point detection with application to smart home time series data [J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(5): 1010 - 1023.

[26]Noor M H M, Salcic Z, Wang K I. Adaptive sliding window segmentation for physical activity recognition using a single tri-axial accelerometer[J]. Pervasive and Mobile Computing, 2017, 38: 41 - 59.

[27]Keogh E, Chu S, Hart D, et al. An on-line algorithm for segmenting time series[C]. Proceedings 2001 IEEE International Conference on Data Mining, USA, 2001: 289 - 296.

[28]Van L K, Berlin E, Schiele B. Enabling efficient time series analysis for wearable activity data[J]. IEEE Press, 2009: 392 - 397.

[29]Berlin E, Van L K. An on-line piecewise linear approximation technique for wireless sensor networks[C]. Proceedings 2010 IEEE Conference on Local Computer Networks, USA, 2010: 905 - 912.

[30]Luo L, You S, Xu Y, et al. Improving the integration of piecewise linear representation and weighted support vector machine for stock trading signal prediction[J]. Applied Soft Computing, 2017, 56: 199 - 216.

[31]王 霞, 洪永淼. 基于非参数回归的遗漏变量检验[J]. 管理科学学报, 2016, 19(3): 77 - 91.
Wang Xia, Hong Yongmiao. Nonparametric-regression-based testing for omitted variables[J]. Journal of Management Sciences in China, 2016, 19(3): 77 - 91. (in Chinese)

[32]Zhou B, Ye H, Zhang H, et al. A new qualitative trend analysis algorithm based on global polynomial fit[J]. Aiche Journal, 2017, 63(8): 3374 - 3383.

[33]蒋志强, 田婧雯, 周炜星. 中国股票市场收益率的可预测性研究[J]. 管理科学学报, 2019, 22(4): 92 - 109.

Jiang Zhiqiang, Tian Jingwen, Zhou Weixing. Return predictability in the Chinese stock markets[J]. Journal of Management Sciences in China, 2019, 22(4): 92 – 109. (in Chinese)

[34] Chen K, Wang J. Finding significant qualitative trend combinations for multivariate systems from historical data[J]. Industrial & Engineering Chemistry Research, 2019, 58(11): 4518 – 4533.

[35] Killick R, Fearnhead P, Eckley I A. Optimal detection of changepoints with a linear computational cost[J]. Journal of the American Statistical Association, 2012, 107(500): 1590 – 1598.

[36] Hocking T D, Rigaill G, Fearnhead P, et al. Generalized functional pruning optimal partitioning (GFPOP) for constrained changepoint detection in genomic data[J]. Journal of Statistical Software, 2022, 101 (10): 1 – 31.

[37] 鄢章华, 刘　蕾. 考虑服务水平与动态转移规律的共享单车投放策略研究[J]. 中国管理科学, 2019, 27(9): 195 – 204.
Yan Zhanghua, Liu Lei. Supply optimization for bicycle sharing system considering service level and dynamic transfer[J]. Chinese Journal of Management Science, 2019, 27(9): 195 – 204. (in Chinese)

[38] Wang C, Wang X, Xing L, et al. A fast and accurate reliability approximation method for heterogeneous cold standby sparing systems[J]. Reliability Engineering & System Safety, 2021, 212: 107596.

[39] Hansen P R, Lunde A, Nason J M. The model confidence set[J]. Econometrica, 2011, 79(2) : 453 – 497.

[40] 梁　超, 魏　宇, 马　锋, 等. 投资者关注对中国黄金价格波动率的影响研究[J]. 系统工程理论与实践, 2022, 42(2): 320 – 332.
Liang Chao, Wei Yu, Ma Feng, et al. A study on the impact of investor attention on Chinese gold volatility[J]. Systems Engineering: Theory & Practice, 2022, 42(2): 320 – 332. (in Chinese)

Forecasting method based on segmentation algorithm of discrete data stream

SUN Li-jun, LI Fang-fang, HU Xiang-pei*

School of Economics and Management, Dalian University of Technology, Dalian 116024, China

Abstract: Discrete data stream has the characteristics of high uncertainty, changeable evolution trend and multi-extremum, which makes it difficult to achieve accurate and real-time forecasting. Therefore, a new forecasting method based on the on-line segmentation algorithm of discrete data stream is proposed. The proposed method combines the short-term trend extraction and on-line adaptive detection of segmentation points. Before forecasting, the short-term trend is obtained based on the nonparametric regression model and the improved trend extraction algorithm, then the short-term trend of the forecasted day is mined and analyzed for later on-line forecasting. In the on-line forecasting stage, the adaptive detection of segmentation points based on hypothesis testing is applied to the on-line data stream, which can solve the problem of determining segmentation points. The forecasting model at the segmentation point is then modified based on the short-term trends, which reduces the dependence of the forecasting model on the buffered data. Finally, numerical results illustrate the validity and feasibility of the proposed method.

Key words: discrete data stream forecasting; on-line data stream segmentation; nonparametric regression; trend extraction